基于 MHSA 和句法关系增强的机器阅读理解方法研究

张 虎 [1] 王宇杰 [1] 谭红叶 [1] 李 茹 [1]

摘 要 机器阅读理解 (Machine reading comprehension, MRC) 是自然语言处理领域中一项重要研究任务, 其目标是通过机器理解给定的阅读材料和问题, 最终实现自动答题. 目前联合观点类问题解答和答案依据挖掘的多任务联合学习研究在机器阅读理解应用中受到广泛关注, 它可以同时给出问题答案和支撑答案的相关证据, 然而现有观点类问题的答题方法在答案线索识别上表现还不是太好, 已有答案依据挖掘方法仍不能较好捕获段落中词语之间的依存关系. 基于此, 引入多头自注意力 (Multi-head self-attention, MHSA) 进一步挖掘阅读材料中观点类问题的文字线索, 改进了观点类问题的自动解答方法; 将句法关系融入到图构建过程中, 提出了基于关联要素关系图的多跳推理方法, 实现了答案支撑句挖掘; 通过联合优化两个子任务, 构建了基于多任务联合学习的阅读理解模型. 在 2020 中国 "法研杯"司法人工智能挑战赛 (China AI Law Challenge 2020, CAIL2020) 和 HotpotQA 数据集上的实验结果表明, 本文提出的方法比已有基线模型的效果更好.

关键词 机器阅读理解, 多头自注意力, 句法关系, 多跳推理

引用格式 张虎, 王宇杰, 谭红叶, 李茹. 基于 MHSA 和句法关系增强的机器阅读理解方法研究. 自动化学报, 2022, **48**(11): 2718−2728

DOI 10.16383/j.aas.c200951

Research on Machine Reading Comprehension Method Based on MHSA and Syntactic Relations Enhancement

ZHANG Hu[1] WANG Yu-Jie[1] TAN Hong-Ye[1] LI Ru[1]

Abstract Machine reading comprehension (MRC), which aims to understand the question and the relevant article to answer questions automatically, is an important research task in natural language processing. Recently, the multi-task joint learning research combining opinion question solving and answer evidence mining has attracted much attention. Although methods proposed by such researches always provide both the answer and the relevant evidence simultaneously, neither are the existing methods handling the opinion-type questions good at identifying the clues to the answer, nor are the previous methods mining the answer evidence good at capturing the dependency relationship between words in the given paragraph. Therefore, the method to solve the opinion-type questions has been improved by further exploring the related text clues within the given reading materials through utilizing multi-head self-attention (MHSA); a multi-hop reasoning method realizing the mining of supporting sentences to the answer has been developed by integrating syntactic relation into the construction process of the element graph; a multi-task joint learning model for MRC has been constructed by optimizing the two sub-tasks jointly. Experiments on MRC datasets of CAIL2020 (China AI Law Challenge 2020) and HotpotQA show that the proposed method can provide better results than the existing baseline models.

Key words Machine reading comprehension (MRC), multi-head self-attention (MHSA), syntactic relations, multi-hop reasoning

Citation Zhang Hu, Wang Yu-Jie, Tan Hong-Ye, Li Ru. Research on machine reading comprehension method based on MHSA and syntactic relations enhancement. *Acta Automatica Sinica*, 2022, **48**(11): 2718−2728

机器阅读理解 (Machine reading comprehension, MRC) 是通过计算机理解文章语义并回答相关问题的一项重要研究任务. MRC 研究对提升机器的自然语言理解能力具有重要促进作用, 已受到学术界和工业界的广泛关注. 早期的 MRC 研究主要采用基于人工规则库的方法, 规则库的建立和维护通常需要耗费大量人力, 且难以回答规则以外的问题[1]. 近年来, 随着机器学习, 特别是深度学习的快速发展[2], MRC 的自动答题效果有了明显提升,

收稿日期 2020-11-16 录用日期 2021-04-02
Manuscript received November 16, 2020; accepted April 2, 2021
国家重点研发计划 (2018YFB1005103), 国家自然科学基金 (62176145), 山西省自然科学基金 (201901D111028) 资助
Supported by National Key Research and Development Program of China (2018YFB1005103), National Natural Science Foundation of China (62176145), and Natural Science Foundation of Shanxi Province (201901D111028)
本文责任编委 张家俊
Recommended by Associate Editor ZHANG Jia-Jun
1. 山西大学计算机与信息技术学院 太原 030006
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006

在一些特定任务中 MRC 模型的回答甚至可以媲美人类水平.

随着 BERT (Bidirectional encoder representations from transformers)[3] 等预训练语言模型的出现, 片段抽取式 MRC 任务的实验结果得到了较大提升, 很多模型在 SQuAD (Stanford question answering dataset)[4] 等数据集上已经超越了人类水平. 为了进一步检验模型的推理能力, 现有很多 MRC 数据集加入了观点类问题, 包括 “是/否”和 “不可回答”问题. SQuAD2.0[5] 在 SQuAD 的基础上增加了不可回答问题; CoQA (Conversational question answering)[6] 是一个多轮对话 MRC 数据集, 它的答案形式涉及片段抽取、是/否、不可回答以及自由回答; CJRC (Chinese judicial reading comprehension)[7] 是首个中文法律 MRC 数据集, 问题类型包括片段抽取、是/否与不可回答问题. 然而, 针对观点类问题的 MRC 任务, 现有阅读理解模型仍然不能得到令人满意的结果. 观点类问题的答案往往不在文章中直接出现, 一般需要通过多个句子推理得出. 因此, 对于此类问题, 模型需要综合理解阅读材料后给出观点, 并且如果根据材料语义无法作答, 模型应该将该问题判定为不可回答.

人类在回答阅读理解问题时, 不仅可以给出问题答案, 而且也可以给出支撑答案的依据. 然而, 现有大多数 MRC 模型仅可以给出问题的答案, 无法给出支撑该答案的答案依据, 得到的答案通常缺乏可解释性. 为提高 MRC 模型的可解释性, 美国卡耐基梅隆大学、美国斯坦福大学等机构联合推出了多文档多跳推理数据集 HotpotQA[8], 要求模型在多个文档里寻找答案线索, 给出答案依据, 并通过推理得到答案; 中国 “法研杯”司法人工智能挑战赛 (China AI Law Challenge 2020, CAIL2020) 阅读理解数据集提出了多跳推理任务, 要求 MRC 模型在回答问题的同时给出答案依据, 即参与推理的句子编号. CAIL2020 阅读理解数据集的样例如图 1 所示.

为了同时实现观点类问题作答和答案依据挖掘, 本文提出了一种多任务联合学习模型 (Multi-task joint learning model, MJL-model). 该模型的主要思想是: 首先, 针对观点类问题, 引入多头自注意力 (Multi-head self-attention, MHSA) 机制挖掘文章中观点类问题的文字线索, 然后利用循环卷积神经网络 (Recurrent convolutional neural network, RCNN)[9] 对观点类问题进行分类求解; 其次, 针对答案依据挖掘任务, 利用词法与句法分析工具识别文章中各句子中的关键要素以及句法关系, 利用要素间的依存句法关系以及其他关联关系构建关联要素关系图, 并利用动态融合图网络 (Dynamically fused graph network, DFGN)[10] 在关系图上挖掘当前问题的答案依据, 增强答案的可解释性; 最后, 通过参数共享与联合损失优化, 将两个任务进行联合优化学习, 实现观点类问题的解答以及答案依据的挖掘. 本文在 CAIL2020 与 HotpotQA 阅读理解数据集上进行了实验, 分析了中英文数据集的差异, 证明了该方法的有效性.

[0]经审理查明: 2011年3月25日, [1]原告购买了淮安禧徕乐投资发展有限公司开发的、座落于本市清浦区枚皋路8号淮安禧徕乐国际商贸城第1幢西B2105号房屋. [2]同日, [3]原、被告（被告当时名称为淮安禧徕乐国际商贸城经营管理有限公司, [4]后变更为现名称）签订《统一经营管理合同》1份, [5]约定原告将购得的上述商铺委托被告统一经营管理, [6]期限5年, [7]自商城开业之日起计算(最迟不晚于2011年12月31日); [8]在商城开业后的前3年为商城培育期, [9]原告同意在此期间内被告无需支付该商铺的租金, [10]第4年至第5年, [11]被告应每年支付原告一次租金为10288元, [12]支付日期为每年的3月1日至3月15日.
Question: 签订合同中是否约定期限?
Answer: yes
Supporting facts: 4, 6

图 1　CAIL2020 阅读理解数据集样例

Fig. 1　Sample of CAIL2020 MRC dataset

本文的主要贡献有以下几点:

1) 提出句法关系增强的关联要素关系图构建方法, 建立基于 DFGN 的答案依据挖掘模型;

2) 针对观点类问题解答和答案依据挖掘任务, 提出多任务联合学习的阅读理解模型;

3) 同时在 CAIL2020 与 HotpotQA 阅读理解数据集上进行了多项对比实验, 验证了所提模型的有效性和通用性.

1　相关工作

1.1　机器阅读理解数据集

近几年, 学术界和工业界提出了多个大规模 MRC 数据集, 促进了 MRC 的发展. RACE (Reading comprehension dataset from examinations)[11] 是美国卡耐基梅隆大学在 2017 年推出的大规模 MRC 数据集, 数据来源为中国中学生的英语考试, 包含了 28 000 篇文章和近 10 万个多项选择题. SQuAD 数据集由斯坦福大学于 2016 年推出, 主要来源于 536 篇维基百科文章, 包含了 10 万多个片段抽取式问题. 2018 年推出的 SQuAD2.0 进一步加入了大量 “无法回答”类问题, 问题数量达到了 15 万个, 答题难度相比 SQuAD有了明显提升. 2017 年, 百度公司基于百度搜索和百度知道数据开放了中文

MRC 数据集 DuReader[12], 该数据集共包含 20 万个问题和 100 万篇相关文档, 问题类型包括自由回答类与 “是/否”类. 2018 年美国卡耐基梅隆大学、美国斯坦福大学等机构基于维基百科数据共同推出了多文档多跳推理 MRC 数据集 HotpotQA, 共包含 11 万个问题, 要求模型答题时能够同时给出答案和答案依据.

1.2 机器阅读理解模型

受到大规模开放阅读理解数据集的驱动, 相关学者对阅读理解模型开展了广泛研究, 在模型设计和训练方法等方面进行了深入探索.

在 BERT 等预训练语言模型提出之前, 最优的 MRC 模型主要探索不同注意力机制的应用. Attentive Reader[13] 首次将注意力机制应用到阅读理解任务中, 它使用双向长短时记忆网络 (Bi-directional long short-term memory, BiLSTM) 对文章和问题进行编码, 计算从问题到文章的注意力. BiDAF (Bidirectional attention flow)[14]将 MRC 模型划分为编码层、交互层和输出层, 建立了文章和问题的交互注意力机制. R-NET[15] 改进了循环神经网络 (Recurrent neural networks, RNN) 在阅读理解任务中的应用, 将注意力机制融入到 RNN, 并通过门控机制动态控制信息的取舍. QANet[16] 摒弃了 RNN 复杂的递归结构, 只使用卷积神经网络和自注意力机制完成编码工作, 提高了模型的速度和准确率.

目前, 预训练语言模型已成为一种新的自然语言处理 (Natural language processing, NLP) 范式, 其主要使用大规模文本语料库进行预训练, 并用特定任务的小数据对模型进行微调, 推动了 MRC 研究的快速发展. Google 于 2018 年推出了 BERT 预训练语言模型, 该模型基于 Transformer 编码器, 引入掩码语言模型 (Masked language model, MLM) 和下一句预测 (Next sentence prediction, NSP) 任务. 随后, 2019 年 Facebook 在 BERT 的基础上提出了 RoBERTa (Robustly optimized BERT approach)[17] 模型, 在预处理阶段采用动态掩码取代了静态掩码, 同时还去掉了 NSP 任务. 显然, 预训练语言模型在提高 NLP 相关任务效果的同时, 也增加了模型参数和训练时长. 针对这些问题, Google 又在 BERT 的基础上提出了 ALBERT (A lite BERT)[18] 模型, 其使用词向量因式分解和跨层参数共享的方法减少了模型的参数量, 同时通过引入句子顺序预测 (Sentence order prediction, SOP) 任务进一步改进了 BERT 模型. 2019 年, 百度推出了中文预训练语言模型 ERNIE (Enhanced representation through knowledge integration)[19], 它通过对词语、实体等语义单元进行掩码 (MASK), 使得模型可以学习到潜在的知识和语义依赖关系, 提高了模型的泛化能力, ERNIE 在中文任务中全面超越了 BERT模型. 随后, 哈尔滨工业大学讯飞联合实验室发布了中文 RoBERTa_wwm_ext[20] 模型, 它将整词掩码 (Whole word masking, WWM) 应用到中文 BERT 模型中, 在多个中文任务中得到了更好的实验结果.

1.3 多跳阅读理解模型

多跳推理要求模型在多个文档中寻找线索并推理出答案, 已成为 MRC 任务中的研究热点, 相关研究人员针对该任务已开展了大量深入研究. CogQA (Cognitive graph question answering)[21] 建立了一种认知图谱问答模型, 它设计了两个系统来维护一张认知图谱, 系统 1 遍历文档, 抽取与问题相关的实体来扩展认知图谱, 系统 2 利用图注意力网络 (Graph attention network, GAT)在构建的认知图谱上进行推理, 并回答问题. DFGN构造了一个动态实体图并通过 GAT 在实体图上进行推理. 同时, 设计了一个融合模块来提高实体图和文章之间的交互性. HDE (Heterogeneous document-entity)[22] 通过互注意力机制学习候选答案、问题、文档以及实体之间的关系, 同时利用这些关系构建了一个异构图, 并通过图卷积神经网络 (Graph convolutional network, GCN) 在异构图上进行推理, 寻找答案证据. QFE (Query focused extractor)[23] 将片段抽取任务与多跳推理任务进行联合学习, 使用 RNN 来依次提取答案支撑句. SAE (Select, answer and explain)[24] 设计了一个筛选模块来过滤文档中与问题无关的句子, 并将片段抽取与多跳推理两个任务进行联合优化, 在多跳推理任务中利用文档句子之间的关系构造关系图, 进而利用 GCN在关系图上进行推理.

2 模型

本文提出的 MJL-model 模型将阅读理解中的片段抽取问题、观点类问题以及答案依据挖掘任务进行联合优化学习, 形成了一个端到端的多任务阅读理解模型. 模型结构如图 2 所示, 主要包括编码层、问题解答层、多跳推理层、预测层. 在问题解答层, 基于 MHSA 及 RCNN 实现了对观点类问题的分类解答; 在多跳推理层, 利用词法和句法分析工具识别文章各句子中的人名、地点、时间、组织机构、名词等关键要素以及要素间的依存句法关系, 利用

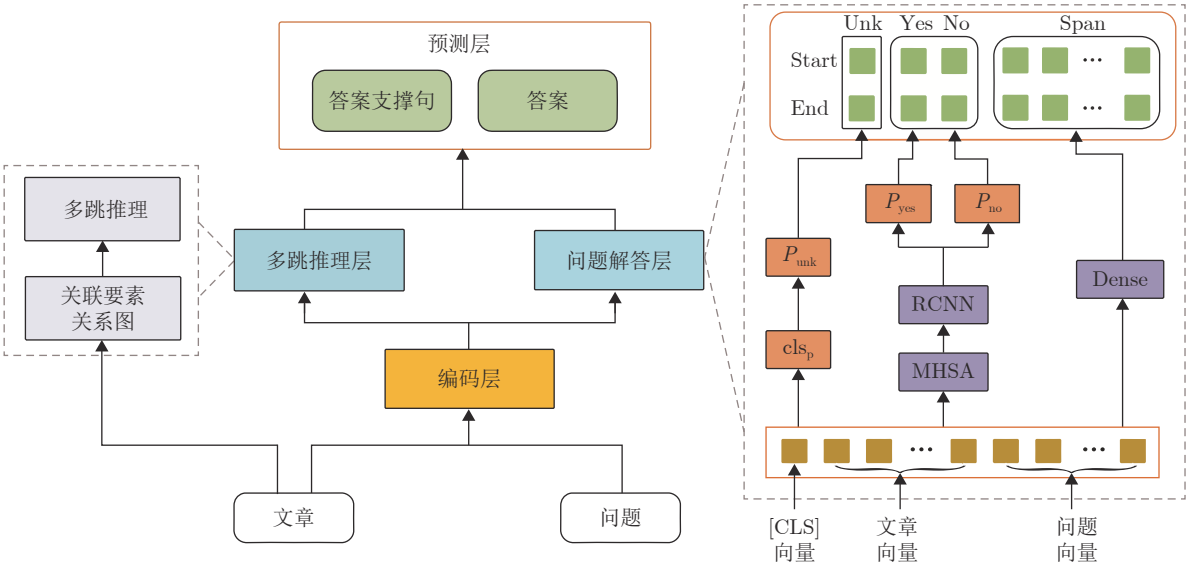


图 2　MJL-model 模型结构

Fig. 2　Model architecture of MJL-model

要素之间的关联关系以及句法关系建立关联要素关系图, 并基于关联要素关系图在 DFGN 模型上进行答案依据挖掘.

2.1　编码层

编码层将文章和问题的每个字或词映射到一个高维的向量空间, 获得每个字或者词的向量表示. 本文使用 RoBERTa_wwm_ext($l=12$,$d=768$) 模型来获得文章 P 和问题 O 的向量化表示, l 代表隐藏层数, d 代表隐藏层大小. 具体如式 (1) 和式 (2) 所示

$$input = [\text{CLS}] + P + [\text{SEP}] + O + [\text{SEP}] \tag{1}$$

$$\boldsymbol{x} = \text{RoBERTa\_wwm\_ext}\,(input) \tag{2}$$

其中, $input$ 表示 RoBERTa_wwm_ext 模型的输入, $\boldsymbol{x}$ 表示文章和问题的 12 层向量表示, 本文使用最后 4 层作为文章和问题的向量表示 $\boldsymbol{u}$, 如式 (3) 和式 (4) 所示

$$\boldsymbol{x} = \left\{\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^l\right\}, \quad l=12 \tag{3}$$

$$\boldsymbol{u} = \text{Concat}\left([\boldsymbol{x}]_{l-3}^{l}\right) \tag{4}$$

2.2　问题解答层

本文将问题类型分为片段抽取 (Span) 类和观点类问题, 其中观点类问题分为是/否 (Yes/No) 类、不可回答 (Unknown) 类. Span 类问题的答案为文章中的一个片段, Yes/No 类问题的答案是 yes 或 no, Unknown 类问题的答案是 unknown. 针对各个类型的问题, 本文采用了不同的处理方法.

1) Yes/No 类

针对 Yes/No 类问题, 模型需要根据文章来回答问题的是否观点, 它的答案不在文章中直接出现, 而需要通过多个句子推理得到. 本文通过引入 MHSA 进一步挖掘文章中 Yes/No 类问题的文字线索, 然后利用 RCNN 实现对该类型问题的分类解答. MHSA 定义为

$$\text{Attention}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

$$head_i = \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{6}$$

$$Multihead = \text{Concat}\,(head_1, \cdots, head_h) \tag{7}$$

其中, $\boldsymbol{Q}\in\mathbf{R}^{n\times d_k}$, $\boldsymbol{K}\in\mathbf{R}^{n\times d_k}$, $\boldsymbol{V}\in\mathbf{R}^{n\times d_k}$, $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 为 $\boldsymbol{u}$ 分别通过 $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$, $\boldsymbol{W}_i^V$ 经过线性变化得到, $\boldsymbol{W}_i^Q\in\mathbf{R}^{4d\times d_k}$, $\boldsymbol{W}_i^K\in\mathbf{R}^{4d\times d_k}$, $\boldsymbol{W}_i^V\in\mathbf{R}^{4d\times d_k}$.

具体而言, 本文将编码层得到的问题与文章的字符向量 $\boldsymbol{u}$ 输入 MHSA 得到新的字符向量表示 $\boldsymbol{u}'$

$$\boldsymbol{u}' = \text{MHSA}\,(\boldsymbol{u}) \tag{8}$$

然后通过 RCNN 和全连接 (Dense) 层进行二分类, 得到问题答案是 yes/no 的概率 $p^{\text{yes}}/p^{\text{no}}$, 具体计算如式 (9) ~ (13) 所示

$$\boldsymbol{u}'' = \text{BiLSTM}\,(\boldsymbol{u}') \tag{9}$$

$$\boldsymbol{y} = \tanh\,(\text{Concat}\,[\boldsymbol{u}'', \boldsymbol{u}']) \tag{10}$$

$$\tilde{\boldsymbol{y}} = \text{MaxPooling}\,(\boldsymbol{y}) \tag{11}$$

$$p^{\text{yn}} = \text{Dense}\,(\tilde{\boldsymbol{y}}) \tag{12}$$

$$\{p^{\text{yes}}, p^{\text{no}}\} = p^{\text{yn}} \tag{13}$$

2) Unknown 类

在观点类问题中, 有些问题仅仅根据文章是无法得到答案的. 对于此类问题, 模型应该拒绝回答. 针对此类问题, 本文用 [CLS] 位置在编码层中得到的向量 $\boldsymbol{c}$ 来表示当前输入的文章和问题, 然后输入一个 $\boldsymbol{W}^c \in \mathbf{R}^{4d\times 1}$ 的 Dense 层, 得到答案是 unknown 的概率 p^{unknown}, 具体计算如 (14) 和式 (15) 所示

$$p^c = \text{Dense}\,(\boldsymbol{c}) \tag{14}$$

$$\left\{p^{\text{unknown}}\right\} = p^c \tag{15}$$

3) Span 类

针对 Span 类问题, 由于它的答案是文章中的一个片段, 模型需要根据问题在文章中标注出正确答案的开始位置和结束位置. 本文通过编码层得到问题及文章每个字符的向量化表示 $\boldsymbol{u}$, 其中文章 P 中 n 个字符的编码为$[\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n]$, $\boldsymbol{u}_i \in \mathbf{R}^{4d}$, 然后在编码层后添加一个 $\boldsymbol{W}^s \in \mathbf{R}^{4d\times 1}$ 的 Dense 层, 获得分数 s, 使用分数 s 来表示每个位置的开始概率 p^s, 具体计算如式 (16) 和式 (17) 所示

$$s = \text{Dense}\,(\boldsymbol{u}) \tag{16}$$

$$p^s = s = [p_1^s, p_2^s, \cdots, p_n^s] \tag{17}$$

同理, 加入另一个 $\boldsymbol{W}^e \in \mathbf{R}^{4d\times 1}$ 的 Dense 层, 获得分数 e, 使用分数 e 来表示每个位置的结束概率 p^e, 如式 (18) 和式 (19) 所示

$$e = \text{Dense}(\boldsymbol{u}) \tag{18}$$

$$p^e = e = [p_1^e, p_2^e, \cdots, p_n^e] \tag{19}$$

2.3 多跳推理层

本文在关联要素关系图上基于 DFGN 模型进行多跳推理, 检索答案依据. 多跳推理层结构如图 3 所示, 主要包括关联要素关系图构建和多跳推理两部分.

在关联要素关系图中, 颜色相同的要素代表它们位于同一句子, 左边关系图考虑了位于同一句子中的要素以及不同句子中的相同要素, 右边关系图考虑了存在句法关系的要素以及相似度大于 η 的要素, 其中 $\eta = 0.90$, 不同类型线条表示了构图过程中不同关系的连边.

1) 关联要素关系图构建

对于 CAIL2020 中文数据集, 本文使用百度开源的 LAC[1] 工具从文章中识别时间、地点、人名、组织、名词、专有名词、数量词等关键要素. 关联要素关系图利用各要素之间的关系进行连边, 在构造关系图时, 本文采用了以下规则: a) 同一句子中的所有要素进行连边; b) 不同句子中的相同要素进行连边; c) 为了增强不同句子之间的要素联系, 本文计算了不同句子中各要素之间的相似度. 首先, 利用 BERT 得到要素对应的词向量, 然后利用余弦相似度计算两个要素之间的相似度, 如果该相似度大于 η, 则对两个要素连边; d) 若不同句子间的两个要素存在句法关系, 则连接两个要素. 首先, 将文章根据句号、问号、感叹号等标点符号进行分割得到片段, 然后使用 DDParser[2] 得到该片段的依存句法关系, 如果两个要素之间存在句法关系, 则连接两个要素.

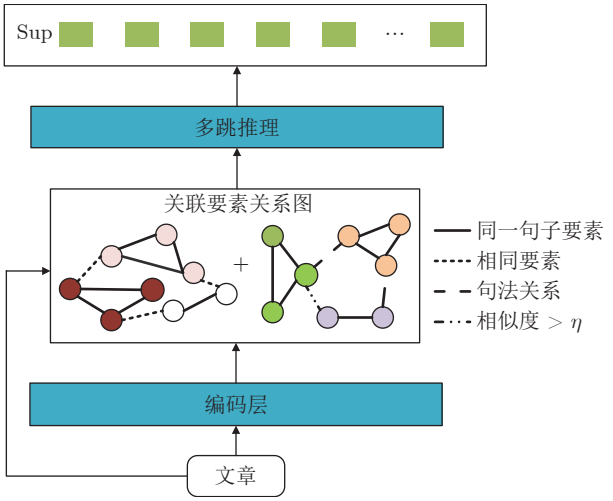


图 3 多跳推理层结构图

Fig. 3 Model architecture of multi-hop reasoning layer

对于 HotpotQA 数据集, 本文使用了 spaCy[3] 从文章中识别时间、地点、人名、名词等关键要素及要素间的依存句法关系.

2) 多跳推理

本文基于已构造的关联要素关系图和 DFGN 进行多跳推理, 具体过程如下:

步骤 1. 本文在数据预处理阶段构建了要素位置矩阵 $\boldsymbol{M}$ 与句子位置矩阵 $\boldsymbol{B}$, $\boldsymbol{M}$ 记录了每个要素在 $input$ 中的相应位置, $\boldsymbol{B}$ 记录了每个句子在 $input$ 中的相应位置, $\boldsymbol{M}$ 与 $\boldsymbol{B}$ 中的元素为 0 或 1.

其中, $\boldsymbol{M}$ 为一个 $w \times g$ 的矩阵, w 表示文章中的要素个数, g 表示 $input$ 的长度, 对于任意要素 i 在 $input$ 中的位置为 $s_i \sim e_i$, 则 $\boldsymbol{M}_{i,\,s_i} \sim \boldsymbol{M}_{i,\,e_i}$ 的值为 1, $\boldsymbol{M}_i$ 中的其余值为 0. $\boldsymbol{B}$ 为一个 $r \times g$ 的矩阵, r 表示文章中的句子个数, 对于任意句子 k 在 $input$ 中的位置为 $s_k \sim e_k$, 则 $\boldsymbol{B}_{k,\,s_k} \sim \boldsymbol{B}_{k,\,e_k}$ 的值为 1, $\boldsymbol{B}_k$ 中的其余值为 0.

[1] https://github.com/baidu/lac
[2] https://github.com/baidu/DDParser
[3] https://github.com/explosion/spaCy

步骤 2. 通过要素位置矩阵 $\boldsymbol{M}$ 得到任意要素 i 在 $input$ 中的相应位置 $s_i \sim e_i$, 在编码层得到了 $input$ 中每个字符的字向量表示 $\boldsymbol{u}$, 则要素 i 对应的字符字向量为 $\boldsymbol{v} = [\boldsymbol{u}_{s_i}, \boldsymbol{u}_{s_i+1}, \cdots, \boldsymbol{u}_{e_i}]$. 本文通过式 (20) 得到要素的词向量 $\boldsymbol{h}$, 初始化关联要素关系图中的要素特征表示.

$$\boldsymbol{h} = \text{MeanPooling}\left(\boldsymbol{v}\right) \tag{20}$$

步骤 3. 通过 MeanPooling 得到问题句向量 $\tilde{\boldsymbol{q}}$, 然后计算关系图中每个要素关于问题的相关度分数 $m = [m_1, m_2, \cdots, m_w]$, 然后通过式 (23) 得到各个要素关于问题的特征表示 $\boldsymbol{h}'$, 使模型在推理过程中更加关注与问题相关的要素.

$$\tilde{\boldsymbol{q}} = \text{MeanPooling}\left(\boldsymbol{q}\right) \tag{21}$$

$$m = \text{Sigmod}\left(\frac{\tilde{\boldsymbol{q}}\boldsymbol{E}\boldsymbol{h}}{\sqrt{\tilde{d}}}\right) \tag{22}$$

$$\boldsymbol{h}' = [m_1\boldsymbol{h}_1, m_2\boldsymbol{h}_2, \cdots, m_w\boldsymbol{h}_w] \tag{23}$$

其中, $\boldsymbol{q}$ 表示通过编码层得到的问题字向量, $\boldsymbol{E}$ 是一个线性变化矩阵.

步骤 4. 基于关联要素关系图进行多跳推理. 首先, 从问题中的某个要素开始推理, 关注在关联要素关系图上与该要素有连边的其他要素. 然后通过计算它们之间的注意力分数, 更新要素的特征表示. 假设对于任意要素 i, 其相邻要素为 N_i, 则要素 i 的注意力权重由式 (24) 和式 (25) 得出

$$e_{ij} = \boldsymbol{A}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}'_i||\boldsymbol{W}\boldsymbol{h}'_j], \quad j \in N_i \tag{24}$$

$$a_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(e_{ij}\right)\right)}{\sum\limits_{k\in N_i}\exp\left(\text{LeakyReLu}\left(e_{ik}\right)\right)} \tag{25}$$

其中, $\boldsymbol{W} \in \mathbf{R}^{F'\times F}$, $\boldsymbol{A} \in \mathbf{R}^{2F\times 1}$, $\boldsymbol{W}$, $\boldsymbol{A}$ 为两个可训练的线性变换矩阵, e_{ij} 表示两个要素之间的相关度分数, a_{ij} 表示要素 i 相对于其相邻要素的注意力权重系数.

最后, 通过式 (26) 计算出要素 i 最终的特征表示 $\tilde{\boldsymbol{h}}_i$

$$\tilde{\boldsymbol{h}}_i = \text{ReLU}\left(\sum_{j\in N_i} a_{j,i}\boldsymbol{W}^h\boldsymbol{h}'_j\right) \tag{26}$$

步骤 5. 每完成一次推理, 使用 Bi-Directional Attention 更新问题的向量表示, 然后通过步骤 3 计算关联要素关系图每个要素关于当前问题向量的相关度分数 m, 并根据 m 去更新关系图的要素特征表示.

最后, 不断重复上述过程更新关联要素关系图各要素的特征表示.

2.4　预测层

预测层基于编码层、问题解答层以及多跳推理层实现了对 Span 类、观点类问题以及答案依据挖掘任务的解答.

1) Span 及观点类问题解答

本文在问题解答层得到了观点类问题的答案概率, 然后将这些答案概率作为 Span 类问题中的答案开始及结束位置概率加入到 Span 类问题中, 与 Span 类问题一起解答. 如式 (27) 和式 (28) 所示, 其中 p^{start}, p^{end} 分别表示每个位置作为答案开始位置和结束位置的概率, n 表示文章长度.

$$p^{\text{start}} = \left\{p^{\text{unknown}}, p^{\text{yes}}, p^{\text{no}}, p_1^s, \cdots, p_n^s\right\} \tag{27}$$

$$p^{\text{end}} = \left\{p^{\text{unknown}}, p^{\text{yes}}, p^{\text{no}}, p_1^e, \cdots, p_n^e\right\} \tag{28}$$

对于 Span 类问题, 由于它的答案是文章中的一个片段, 答案位置需要满足 $1 \le b \le f$ 和 $f \le n$, 其中 b 表示答案的开始位置, f 表示答案的结束位置, 本文将开始位置和结束位置的概率之和作为答案概率. 在 Span 类问题中, 符合上述条件的答案一般有多个, 本文从多个答案中选择概率最大的作为 Span 类问题的答案. 同理, 对于观点类问题也需要计算答案概率. 本文将观点类问题的概率的 2 倍作为答案概率. 最后从多个答案中选择答案概率最大的作为最终答案, 具体计算如式 (29)~(33) 所示

$$p_{\text{Span}}^{\text{Answer}} = \text{argmax}\left(p_b^{\text{start}} + p_f^{\text{end}}\right), \quad 1 \le b \le f,\ f \le n \tag{29}$$

$$p_{\text{Yes}}^{\text{Answer}} = p^{\text{yes}} \times 2 \tag{30}$$

$$p_{\text{No}}^{\text{Answer}} = p^{\text{no}} \times 2 \tag{31}$$

$$p_{\text{Unknown}}^{\text{Answer}} = p^{\text{unknown}} \times 2 \tag{32}$$

$$p^{\text{Answer}} = \text{argmax}\left(\begin{bmatrix} p_{\text{Span}}^{\text{Answer}}, p_{\text{Yes}}^{\text{Answer}} \\ p_{\text{No}}^{\text{Answer}}, p_{\text{Unknown}}^{\text{Answer}} \end{bmatrix}\right) \tag{33}$$

最后, 当 $p^{\text{Answer}} = p_{\text{Span}}^{\text{Answer}}$, 则模型根据答案的起始位置在文章中截取某一连续片段作为问题答案; 当 $p^{\text{Answer}} = p_{\text{Yes}}^{\text{Answer}}$, 答案为 “yes”; 当 $p^{\text{Answer}} = p_{\text{No}}^{\text{Answer}}$, 答案为 “no”; 当 $p^{\text{Answer}} = p_{\text{Unknown}}^{\text{Answer}}$, 答案为 “unknown”.

2) 答案依据挖掘

本文通过多跳推理得到了关联要素关系图中每个要素的特征表示 $\tilde{\boldsymbol{h}}$, 结合要素位置矩阵 $\boldsymbol{M}$ 得到了要素对应字符的字向量表示, 并进一步通过长短期记忆网络 (Long short-term memory, LSTM) 得到文章 P 的特征表示 $\boldsymbol{z}$. 然后结合句子位置矩阵 $\boldsymbol{B}$, 通过 Mean-Max Pooling 得到文章 P 中 r 个句子的特征表示 $\tilde{\boldsymbol{z}}$, 具体计算如式 (34) 和式 (35) 所示

$$\boldsymbol{z} = \text{LSTM}\left(\left[\boldsymbol{u},\ \boldsymbol{M}\tilde{\boldsymbol{h}}^{\text{T}}\right]\right) \tag{34}$$

$$\tilde{\boldsymbol{z}} = \text{Concat}[\text{MeanPooling}(\boldsymbol{B}\boldsymbol{z}^{\text{T}}),\ \text{MaxPooling}(\boldsymbol{B}\boldsymbol{z}^{\text{T}})] \tag{35}$$

然后通过 Dense 层得到 r 个句子关于问题的相关度分数 t, 使用 t 来表示每个句子对于文章问题的支持率 p^{sup}, 具体如式 (36) 和式 (37) 所示

$$t = \text{Dense}\,(\tilde{\boldsymbol{z}}) \tag{36}$$

$$p^{\text{sup}} = t = \left[p_1^{\text{sup}}, p_2^{\text{sup}}, \cdots, p_r^{\text{sup}}\right] \tag{37}$$

实验选择 $p^{\text{sup}} > 0.53$ 的句子作为支撑问题答案的依据.

3 实验

3.1 数据集

本文分别在 CAIL2020 阅读理解数据集和 HotpotQA 数据集上进行了实验.

CAIL2020 阅读理解数据集包括民事、刑事和行政共 3 类中文裁判文书, 问题类型涉及 Span 类、Yes/No 类以及 Unknown 类, 且每个问题都需要给出答案依据. 具体而言, 对于每个问题, 需要结合案情描述内容给出回答, 同时需要给出支撑答案的依据, 即所有支撑答案的句子编号. 由于目前 CAIL2020 只公布了训练集, 没有公布验证集与测试集, 在实验中, 本文根据各问题类型在整体数据中的比例按照 4 : 1 的比例划分了训练集与测试集.

HotpotQA 数据集与 CAIL2020 司法阅读理解数据集较为相似, 两个数据集的任务形式基本一致. HotpotQA 数据集为每个问题提供了 10 篇文章, 问题类型包括 Span 类和 Yes/No 类, 要求对每个问题给出答案和答案依据. 本文在 HotpotQA Distractor 验证集上进行了实验.

3.2 评价指标

实验所用的评价指标包括 3 个部分, 分别是 Span 类和观点类问题的 F1 值 (Ans_F1)、答案依据挖掘任务的 F1 值 (Sup_F1) 以及两部分的联合 F1 值 ($Joint\_F1$).

Ans_F1 计算过程如式 (38) ~ (40) 所示

$$Precision^{\text{Ans}} = \frac{w_c}{w_p} \tag{38}$$

$$Recall^{\text{Ans}} = \frac{w_c}{w_g} \tag{39}$$

$$Ans\_F1 = \frac{2 \times Precision^{\text{Ans}} \times Recall^{\text{Ans}}}{Precision^{\text{Ans}} + Recall^{\text{Ans}}} \times 100\% \tag{40}$$

其中, w_c 表示预测答案与真实答案中相同的字符数, w_p 表示预测答案所包含的字符数, w_g 表示真实答案所包含的字符数.

Sup_F1 计算过程如式 (41) ~ (43) 所示

$$Precision^{\text{Sup}} = \frac{TP}{TP + FP} \tag{41}$$

$$Recall^{\text{Sup}} = \frac{TP}{TP + FN} \tag{42}$$

$$Sup\_F1 = \frac{2 \times Precision^{\text{Sup}} \times Recall^{\text{Sup}}}{Precision^{\text{Sup}} + Recall^{\text{Sup}}} \times 100\% \tag{43}$$

其中, TP 表示预测答案与真实答案均为支撑句的句子数; FP 表示预测答案是支撑句但真实答案不是支撑句的句子数; FN 表示预测答案不是支撑句但真实答案是支撑句的句子数.

$Joint\_F1$ 的计算过程如式 (44) ~ (46) 所示

$$Precision^{\text{Joint}} = Precision^{\text{Ans}} \times Precision^{\text{Sup}} \tag{44}$$

$$Recall^{\text{Joint}} = Recall^{\text{Ans}} \times Recall^{\text{Sup}} \tag{45}$$

$$Joint\_F1 = \frac{2 \times Precision^{\text{Joint}} \times Recall^{\text{Joint}}}{Precision^{\text{Joint}} + Recall^{\text{Joint}}} \times 100\% \tag{46}$$

3.3 基线模型

实验中采用 5 个模型作为 CAIL2020 数据集的基线模型, 分别为:

1) Baseline_BERT (RoBERTa)[4]: CAIL2020 阅读理解任务提供的基于 BERT 的阅读理解模型;

2) Baseline_DPCNN: 将 MJL-model 模型中的 RCNN 替换为深度金字塔卷积神经网络 (Deep pyramid convolutional neural network, DPCNN)[25];

3) Cola[5] (Single model): CAIL2020 阅读理解挑战赛第 4 名所用模型;

4) DFGN_CAIL: 按照 CAIL2020 的数据格式, 修改了 DFGN 的数据处理部分.

实验中采用 4 个模型作为 HotpotQA 数据集的基线模型, 分别为:

1) Baseline[6]: HotpotQA 阅读理解任务提供的基于 Glove (Global vectors)[26] 的阅读理解模型;

2) QFE: 通过注意力机制和 RNN 进行推理, 并将片段抽取与多跳推理任务进行联合优化;

3) DFGN: 根据实体间的关系构造动态实体

[4] https://github.com/china-ai-law-challenge/CAIL2020/tree/master/ydlj

[5] https://github.com/neng245547874/cail2020-mrc

[6] https://github.com/hotpotqa/hotpot

图, 通过 GAT 在实体图上进行多跳推理;

4) SAE: 利用文档句子间的关系构造关系图, 通过 GCN 在关系图上进行多跳推理.

3.4　实验结果

1) CAIL2020 数据集实验结果

对于 CAIL2020 提供的基线模型, 本文分别采用了 BERT_base 和 RoBERTa_wwm_ext 作为模型的编码器. 各模型均采用了相同的参数设置, 具体为: lr = 0.00002, epoch = 10, dropout = 0.1, batch_size = 6, seq_length = 512, 实验结果如表 1 所示. 由表 1 可以看出, Baseline_RoBERTa 模型的 *Ans_F*1 相比 Baseline_BERT 提高了 1.41 个百分点, *Sup_F*1 提高了 5.37 个百分点, *Joint_F*1 提高了 6.49 个百分点. 因此, 本文提出的方法和采用的基线模型均采用了 RoBERTa_wwm_ext 作为编码器. 不同模型的实验结果显示, 本文提出的 MJL-model 模型在 3 项评价指标上都优于所有基线模型.

表 1　CAIL2020 数据集实验结果 (%)

Table 1　Results on the CAIL2020 dataset (%)

模型	*Ans_F*1	*Sup_F*1	*Joint_F*1
Baseline_BERT	70.40	65.74	49.25
Baseline_RoBERTa	71.81	71.11	55.74
Baseline_DPCNN	77.43	75.07	61.80
Cola	74.63	73.68	59.62
DFGN_CAIL	68.79	72.34	53.82
MJL-model	**78.83**	**75.51**	**62.72**

2) HotpotQA 数据集实验结果

同时, 本文在 HotpotQA Distractor 验证集上进一步验证了提出的方法, 且 MJL-model 模型采用与基线模型 DFGN、SAE 完全相同的 BERT_base_uncase 模型作为编码器.

由表 2 可以看出, 本文提出的 MJL-model 模型的 *Ans_F*1 相比 Baseline 模型提高了 12.64 个百分点, *Sup_F*1 提高了 19.30 个百分点, *Joint_F*1 提高了 22.01 个百分点. MJL-model 3 项评价指标都优于 Baseline、QFE、DFGN, 并且 *Sup_F*1 优于所有基线模型. 不同模型的实验结果表明了本文提出的 MJL-model 模型的有效性.

3) 实验数据分析

通过分析模型的实验结果和所用的两个数据集, 发现 MJL-model 模型在中、英文数据集上的表现存在一些差异, 具体原因包括以下 3 个方面:

a) 数据集存在差异. CAIL2020 数据集按照逗号、分号、句号等将一篇文章划分为不同的句子, 相邻句子存在较强的关联性, 但句子间包括的相同词汇较少; HotpotQA 数据集中的每条句子相对独立, 相邻句子间关联性较弱, 且不同句子间存在较多的相同单词.

表 2　HotpotQA 实验结果 (%)

Table 2　Results on the HotpotQA dataset (%)

模型	*Ans_F*1	*Sup_F*1	*Joint_F*1
Baseline	58.28	66.66	40.86
QFE	68.70	84.70	60.60
DFGN	69.34	82.24	59.86
SAE	**74.81**	85.27	**66.45**
MJL-Model	70.92	**85.96**	62.87

b) 构图上存在差异. 由于 CAIL2020 数据集中不同句子间的相同词汇较少, 利用句法关系来增强不同句子间的词汇联系, 可以进一步帮助模型推理出答案句. HotpotQA 数据集考虑了一般的命名实体和名词性单词, 不同句子间相同实体及单词出现的次数较多, 同时由于每条句子较为独立, 因此只有少数相邻句子间存在句法关系.

c) 句法分析工具存在差异. 中文句法分析工具可以分析普通词汇、命名实体间的句法关系; 英文句法分析工具 SpaCy、Stanford CoreNLP 等在进行句法分析时是以单词粒度进行的, 不能将命名实体作为一个整体去考虑.

因此, 本文提出的模型在中文数据集上能够扩充更多的节点关系, 实验结果也比英文数据集的结果更好.

3.5　消融实验

为了进一步评估模型各个模块的贡献, 本文进行了以下消融实验:

1) Question_answering: 将片段抽取和观点类问题作为单任务进行实验;

2) Answer_evidence: 将答案依据挖掘任务作为单任务进行实验;

3) –MHSA: 去掉问题解答层中的多头自注意力;

4) –RCNN: 去掉问题解答层中的循环卷积神经网络;

5) –Syntax & Similarity: 在构建要素关系图时, 去掉要素之间的句法以及相似度关系.

具体消融实验结果如表 3 所示.

表 3 实验结果显示, Question_answering 的 *Ans_F*1 与 Answer_evidence 的 *Sup_F*1 相比 MJL-model 都下降了 2 个多百分点, 证明了多任务

表 3 消融实验结果 (%)

Table 3 Results of ablation experiments (%)

模型	*Ans_F*1	*Sup_F*1	*Joint_F*1
MJL-model	**78.83**	**75.51**	**62.72**
Question_answering	76.36	—	—
Answer_evidence	—	73.42	—
–MHSA	76.28	75.11	61.16
–RCNN	75.96	75.05	60.96
–Syntax & Similarity	77.61	74.39	60.80

联合优化的有效性; 针对观点问题解答层, 去掉 MHSA 后 *Ans_F*1 下降了 2.55 个百分点, 去掉 RCNN 后 *Ans_F*1 下降了 2.87 个百分点, *Sup_F*1 及 *Joint_F*1 也都有明显下降; 针对关联要素关系图, 去掉要素之间的句法关系以及相似度关系, *Sup_F*1 下降了 1.12 个百分点, *Ans_F*1 下降了 1.22 个百分点, *Joint_F*1 下降了 1.92 个百分点. 通过对消融实验结果的分析, 证明了本文所提方法的有效性.

3.6 模型有效性分析

为了进一步验证 MHSA 机制和句法关系对模型结果的影响, 本文对两个样例的关键过程进行了可视化展示, 具体样例如图 4 ~ 6 所示.

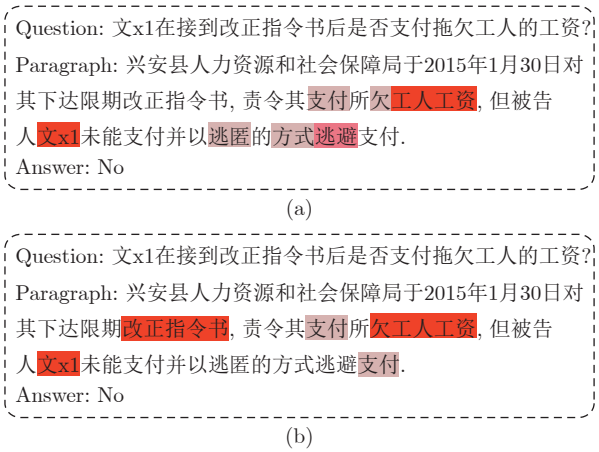


图 4 注意力可视化样例

Fig. 4 Sample of attention visualization

1) 图 4 呈现了实验数据中某问题对应语句片段的注意力可视化样例, 其中颜色越深, 代表它的注意力权重越高, 对于模型正确作答越重要. 图 4(a) 为引入 MHSA 机制的示例, 图 4(b) 为去掉 MHSA 机制的示例.

显然, 引入 MHSA 机制后, 模型不仅关注问题中出现的词汇, 而且也能捕获带有观点类文字线索的词汇, 例如 “逃匿”、“逃避”; 而去掉 MHSA 机制后, 模型仅关注 “文 x1”、“支付” 等在问题中出现

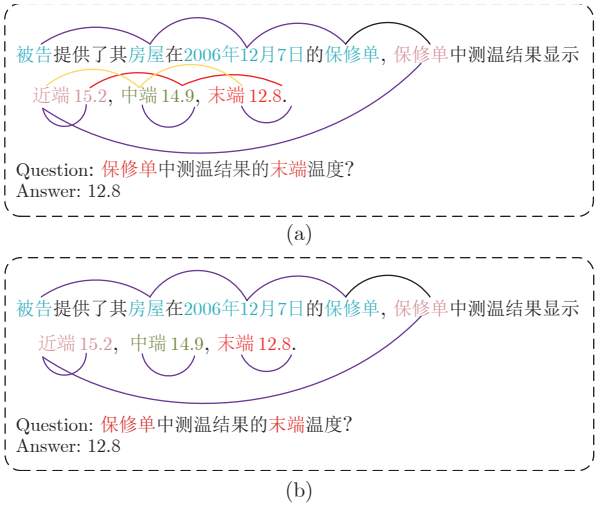


图 5 关联要素关系图样例

Fig. 5 Sample of related element graph

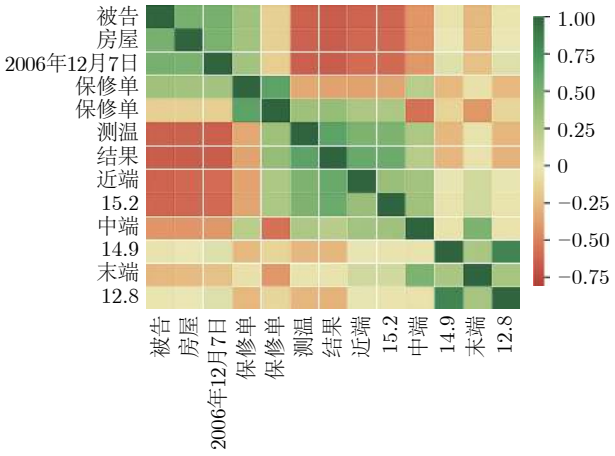


图 6 多跳推理注意力可视化样例图

Fig. 6 Visible sample of multi-hop reasoning attention

的词汇, 对观点类文字线索的关注较少. 因此, 引入 MHSA 机制可以使模型更好地回答观点类问题.

2) 图 5 展示出实验数据集中一个真实语句片段生成的关联要素关系图样例, 图 5(a) 为融入依存句法关系和要素相似度的示例, 图 5(b) 为 DFGN 生成的句子示例.

图 5(a) 根据本文提出的关联规则将各要素进行连接, 可得到 “15.2”、“14.9”、“12.8”在句法上存在并列关系, “近端”、“中端”和 “末端”间的相似度大于 η, 根据关系图构建规则可将这 3 个要素连接. 基于该图, 模型可从问题要素 “保修单”出发, 得到 “保修单–末端–12.8”线索关系. 图 5(b) 仅考虑了同一句子中的所有要素以及不同句子中的相同要素, 缺乏能够支撑问题与答案的线索关系. 同时, 为了更直观地展示推理过程中要素之间的注意力关

系, 进一步输出了上述样例的多跳推理注意力可视化, 如图 6 所示.

从图 6 可以看出, “保修单”与 “近端”、“中端”、“15.2” 等要素具有较强的关联性, “近端”与 “15.2”、“中端”、“末端” 等要素紧密关联, “中端”与 “近端”、“15.2”、“末端” 等要素有紧密联系, “末端”与 “中端”、“14.9”、“12.8” 等要素关联性较强. 显然, 可以进一步建立 “保修单”与 “末端”和 “12.8” 的关联关系. 因此, 本文提出的关联要素关系图能得到更有效的实验结果.

4 结束语

本文针对阅读理解任务中的观点类问题以及答案依据挖掘展开研究, 提出了一种基于 MHSA 与句法关系增强的多任务阅读理解模型. 通过引入 MHSA 和 RCNN, 改进了观点类问题的解答方法; 利用句法关系与其他要素关系构建关联要素关系图, 并基于关联要素关系图进行多跳推理, 优化了答案依据挖掘模型; 最后将两个任务进行联合优化学习, 建立了基于多任务联合学习的阅读理解模型. 在 CAIL2020 阅读理解数据集和 HotpotQA 数据集上的成功应用, 验证了所提方法的有效性.

在观点类问题中, 仅通过 MHSA 机制挖掘文章中观点类问题的文字线索可能还不够充分. 在未来工作中, 将尝试利用图神经网络来进一步挖掘文章中观点类文字线索; 答案依据挖掘对于阅读理解的可解释性具有重要意义, 下一步将引入一些外部知识库[27] 和其他推理方法来探索更有效的答案依据挖掘方法.

References

1 Zeng Shuai, Wang Shuai, Yuan Yong, Ni Xiao-Chun, Ouyang Yong-Ji. Towards knowledge automation: A survey on question answering systems. *Acta Automatica Sinica*, 2017, **43**(9): 1491−1508
(曾帅, 王帅, 袁勇, 倪晓春, 欧阳永基. 面向知识自动化的自动问答研究进展. 自动化学报, 2017, **43**(9): 1491−1508)

2 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

3 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: ACL, 2019. 4171−4186

4 Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQUAD: 100 000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA: ACL, 2016. 2383−2392

5 Rajpurkar P, Jia R, Liang P. Know what you don′t know: Unanswerable questions for squad. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018.784−789

6 Reddy S, Chen D Q, Manning C D, CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7: 249−266

7 Duan X Y, Wang B X, Wang Z Y, Ma W T, Cui Y M, Wu D Y, et al. CJRC: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. In: Proceedings of the 2019 China National Conference on Chinese Computational Linguistics. Kunming, China: Springer, 2019. 439−451

8 Yang Z L, Qi P, Zhang S Z, Bengio Y, Cohen W W, Salakhutdinov R, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 2369−2380

9 Lai S W, Xu L H, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. In: Proceedings of the 2015 AAAI Conference on Artificial Intelligence. Austin, USA: AAAI, 2015. 2267−2273

10 Xiao Y X, Qu Y R, Qiu L, Zhou H, Li L, Zhang W N, Yu Y. Dynamically fused graph network for multi-hop reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 6140−6150

11 Lai G K, Xie Q Z, Liu H X, Yang Y M, Hovy E. RACE: Large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: ACL, 2017. 785−794

12 He W, Liu K, Liu J, Lv Y J, Zhao S Q, Xiao X Y, et al. Dureader: A Chinese machine reading comprehension dataset from real-world applications. In: Proceedings of the 2018 Workshop on Machine Reading for Question Answering. Melbourne, Australia: ACL, 2018. 37−46

13 Chen D Q, Bolton J, Manning C D. A thorough examination of the CNN/daily mail reading comprehension task. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016. 2358−2376

14 Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.

15 Wang W H, Yang N, Wei F R, Chang B B, Zhou M. Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017. 189−198

16 Yu A W, Dohan D, Luong M T. QANet: Combining local convolution with global self-attention for reading comprehension. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018.

17 Liu Y H, Ott M, Goyal N, Du J F, Joshi M, Chen D Q, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.

18 Lan Z Z, Chen M D, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A lite BERT for self-supervised learning of language representations. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.

19 Sun Y, Wang S H, Li Y K, Feng S K, Chen X Y, Zhang H, et al. ERNIE: Enhanced representation through knowledge integration. arXiv: 1904.09223, 2019.

20 Cui Y M, Che W X, Liu T, Qin B, Yang Z Q. Pre-training with whole word masking for Chinese BERT. IEEE Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504−3514

21 Ding M, Zhou C, Chen Q B, Yang H X, Tang J. Cognitive graph for multi-hop reading comprehension at scale. In: Proceedings of the 57th Annual Meeting of the Association for Compu-

tational Linguistics. Florence, Italy: ACL, 2019. 2694−2703

22 Tu M, Wang G T, Huang J, Tang Y, He X D, Zhou B W. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 2704−2713

23 Nishida K, Nishida K, Nagata M, Otsuka A, Saito I, Asano H, et al. Answering while summarizing: multi-task learning for multi-hop QA with evidence extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 2335−2345

24 Tu M, Huang K, Wang G T, Huang J, He X D, Zhou B W. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In: Proceedings of the 32nd Innovative Applications of Artificial Intelligence Conference. New York, USA: AAAI, 2020. 9073−9080

25 Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017. 562−570

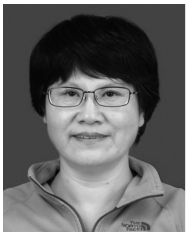
26 Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014. 1532−1543

27 Liu Kang, Zhang Yuan-Zhe, Ji Guo-Liang, Lai Si-Wei, Zhao Jun. Representation learning for question answering over knowledge base: An overview. *Acta Automatica Sinica*, 2016, **42**(6): 807−818
(刘康, 张元哲, 纪国良, 来斯惟, 赵军. 基于表示学习的知识库问答研究进展与展望. 自动化学报, 2016, **42**(6): 807−818)

张 虎 山西大学计算机与信息技术学院副教授. 2014 年于山西大学计算机与信息技术学院获得工学博士学位. 主要研究方向为人工智能与自然语言处理. 本文通信作者.
E-mail: zhanghu@sxu.edu.cn
(**ZHANG Hu** Associate professor at the School of Computer and Information Technology, Shanxi University. He received his Ph.D. degree from the School of Computer and Information Technology, Shanxi University in 2014. His research interest covers artificial intelligence and natural language processing. Corresponding author of this paper.)

王宇杰 山西大学计算机与信息技术学院博士研究生. 主要研究方向为自然语言处理.
E-mail: init_wang@foxmail.com
(**WANG Yu-Jie** Ph.D. candidate at the School of Computer and Information Technology, Shanxi University. His main research interest is natural language processing.)

谭红叶 山西大学计算机与信息技术学院教授. 2008 年于哈尔滨工业大学计算机学院获得博士学位. 主要研究方向为人工智能, 自然语言处理.
E-mail: tanhongye@sxu.edu.cn
(**TAN Hong-Ye** Professor at the School of Computer and Information Technology, Shanxi University. She received her Ph.D. degree from the School of Computer, Harbin Institute of Technology in 2008. Her research interest covers artificial intelligence and natural language processing.)



李 茹 山西大学计算机与信息技术学院教授. 2011 年于山西大学计算机与信息技术学院获得工学博士学位. 主要研究方向为人工智能与自然语言处理. E-mail: liru@sxu.edu.cn
(**LI Ru** Professor at the School of Computer and Information Technology, Shanxi University. She received her Ph.D. degree from the School of Computer and Information Technology, Shanxi University in 2011. Her research interest covers artificial intelligence and natural language processing.)